International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.10, No. 1 (2017), pp.1-8
http://dx.doi.org/10.14257/ijsip.2017.10.1.01

Online Salient Object Segmentation with Local Weight Assignment

Song Gu[1,2], Zheng Ma[2] and Mei Xie[3]

[1]Chengdu Aeronautic Polytechnic
[2]School of Communication and Information Engineering, University of Electronic Science and Technology of China
[3]School of Electronic Engineering, University of Electronic Science and Technology of China
Gusong1215@icloud.com

Abstract

We present a novel on-line algorithm for target segmentation and tracking in video. In our approach, video data is represented by both temporal saliency and spatial one, and segmentation is accomplished by finding the minimum energy label assignment. The pixel-wise weight is assigned for each energy item according to the local information of each feature map. This local weight model enhances the segmentation accuracy. Experiments demonstrates that our approach is effective.

Keywords: *salient object segmentation, local weight assignment, L2ECM*

1. Introduction

Image segmentation aims to group perceptually similar pixels into regions and it is a fundamental problem in computer vision. Video segmentation generalizes this concept to group pixels into spatio temporal regions that exhibit coherence in both appearance and motion. Most previous works have proposed the segmentation solutions based on graph-cut. Different cues have been formalized in these methods, and merged together in the same framework referred as energy function. The label assignment solution is converted into the energy minimization one which can be solved by graph-cut [1-2]. In resolving the energy function, a weight factor is assigned for each energy item to control the proportion in the whole. For example, [2] proposed a Fisher linear discriminant to measure the discriminate performance of each feature map. In [3], a variance ratio measurement is adopted to adaptively adjust the weights of different features. [4] measures the discriminate power by computing the KL distance of histogram. Generally, the weights are motivated by the fact that the segmentation cues with high foreground/background discrimination deserve high weights so that they can make a significant contribution to the segmentation in the next frame. The weight of each energy item is a real value which is computed by the global information of different features in these methods. However, in some cases, each segmentation cue has different discriminate performances in different regions. The local information will obtain better property than the global one in these cases.

Our goal is to address this problem within the same framework. Thanks to the idea in [5], an element uniqueness measurement, described as temporal salient map, is developed by two successive frames in our approach. The spatial salient map is designed in feature space which is different from the method proposed in [5] as well. Both measurements are computed in the same framework, high-dimensional Gaussian filtering, and two segmentation cues in energy function are constructed by both maps respectively. To obtain an accurate pixel-wise segmentation, we choose to integrate both cues into energy minimization framework as a constraint instead of directly outputting it like [5]. An adaptive weight of each energy item is assigned for each pixel according to its discriminate performance within a certain neighborhood.

2. Image Representation

In this paper, we use super-pixels to abstract the image into perceptually uniform regions. Considering the irregular shape of the super-pixel, L2ECM for image representation is adopted. Provided with the raw feature vectors, we can obtain the L2ECM features for each super-pixel as [6]. Note that the covariance matrix is computed in each super-pixel in our approach.

2.1. Feature Abstraction

In this paper, we use an adaptation of SLIC super-pixels [7] to abstract the image into perceptually uniform regions. SLIC super-pixel representation not only reduces computational complexity in later stages of processing, but also makes computation more robust by enforcing consistency inside super-pixels. Considering the irregular shape of the super-pixel, L2ECM for image representation is adopted. Given an image, some raw features are formulated as

$$f(x,y)=[I(x,y),\left|I_x(x,y)\right|,\left|I_y(x,y)\right|,\left|I_{xx}(x,y)\right|,\left|I_{yy}(x,y)\right|]^T \quad (1)$$

where $|\bullet|$ denotes the absolute value, $I(x,y)$ denotes the intensity of a pixel locating (x,y) in the image, $I_x(x,y)$ (resp. $I_{xx}(x,y)$) and $I_y(x,y)$ (resp. $I_{yy}(x,y)$) denote the first (resp. second)-order partial derivative with respect to x and y respectively.

Given a super-pixel *S*, let $\left\{f(x_i,y_i)\in R^d\right\}_{i=1}^{N_s},(x_i,y_i)\in S$ be the feature points inside *S*. The super-pixel is represented by the covariance matrix $C_S \in R^{d\times d}$

$$C_s(i,j)=\frac{1}{N_s}(f(x_i,y_i)-\mu)(f(x_i,y_i)-\mu)^T \quad (2)$$

where μ is the mean of the feature points which is defined as $\mu=\frac{1}{N_s}\sum_{i=1}^{N_s}f(x_i,y_i)$,$C_s(i,j)$ denotes the element at the *i-th* row, *j-th* column of C_S,N_s is the number of pixels inside *S*, *d* is the length of the raw feature $f(x,y)$.

To avoid computing the geodesic distance between covariances that lie on Riemannian manifold, we transform C_S into $\log(C_S)$ that locates in Euclidean space with matrix logarithm operation and construct the L2ECM feature for each super-pixel by performing half-vectorization of $\log(C_S)$ which is proposed in [6]. L2ECM feature for a super-pixel *i* in current image is a $\frac{d(d+1)}{2}$ length vector which is described as f_i^t. L2ECM feature is adopted as a feature abstraction algorithm in our approach, because it has the following advantages:

(1) The theoretical foundation of L2ECM is the Log-Euclidean framework, which endows the commutative lie group formed by the SPD (Symmetric and Positive Definite) matrices with a liner space structure. This enables the common Euclidean operations of covariance matrices in the logarithmic domain while preserving their geometric structure.

(2) The dimension of L2ECM feature vector is only related with the dimension of raw feature vectors regardless of the size and shape of estimated region such as super-pixel, which implies a certain scale and rotation invariance over the regions in different images. This kind of feature is better applicable to region-based algorithm than features that other region-based video segmentation methods used.

(3) The noise-corrupting individual samples are largely filtered out with the average filter during covariance computation.

2.2. Temporal Salient Map

Let $\left\{f_i^t\right\},(i=1,2,3\cdots M)$ be the feature set of each super-pixel in current image f^t. M is the number of the super-pixels. Assumed that the object is segmented well in previous image f^{t-1}. The feature set in previous image which belongs to the object is defined as $\left\{f_i^{t-1}\right\},(i=1,2,3\cdots N)$ where N is the number of the super-pixels which belongs to the object in previous image. We directly compute the temporal saliency of each super-pixel in current image as

$$T_i=\sum_{j=1}^{N}\frac{1}{d(f_i^t,f_j^{t-1})}w(p_i^t,p_j^{t-1}),i=1,2,3\cdots M \tag{3}$$

where $w(p_i^t,p_j^{t-1})=\frac{1}{Z_i}\exp(-\frac{1}{2\sigma_p^2}\left\|p_i^t-p_j^{t-1}\right\|^2)$ is a Gaussian weight function, p_i^t and p_j^{t-1} represent locations of the corresponding super-pixels in f^t and f^{t-1} respectively. $d(f_i^t,f_j^{t-1})$ is a distance measurement between f^t and f^{t-1}. In addition, we find that the cosine distance outperforms the Euclidean one in L2ECM space. Z_i is the normalization factor ensuring $\sum_{j=1}^{N}w(p_i^t,p_j^{t-1})=1$, and σ_p controls the range of the temporal salient operator.

The interpretation of Equation 3 is intuitive. Given a super-pixel *i* in f^t, the feature and the location of the give super-pixel are compared to all the super-pixels belonging to the object in f^{t-1}. $d(f_i^t,f_j^{t-1})$ formulates the similarity of both super-pixels f_i^t and f_j^{t-1}. $w(p_i^t,p_j^{t-1})$ is related to the spatial distance between both super-pixels. A super-pixel in f^t is more likely to be regarded as the object, both in feature space and in spatial space, under the condition that it is similar to the super-pixels in f^{t-1}. Overall, the temporal saliency of each super-pixel, T_i, formulates the probability of each super-pixel which belongs to the object in current frame ac- cording to previous segmentation.

Note that although Equation 3 has the same structure as the 'element uniqueness' in [5], both equations show different meaning. The temporal saliency in our approach formulates the similarity of the object in two successive frames. The element uniqueness in [5] analyzes the difference between the object and the background in an individual image. The subscript *i* and *j* in Equation 3 represent the super-pixel in current frame and in previous frame respectively.

2.3. Spatial Salient Map

Temporal saliency exhibits the object similarity between two consecutive frames. However, lighting variation and noise might be incorrectly assigned to dynamic objects.

Spatial saliency is always adopted to reduce errors resulting from temporal saliency. The difference between temporal saliency and spatial saliency is that the former is measured in some successive frames, while the latter is computed in an individual image. In our approach, we estimate the distribution of each super-pixel in current frame as the spatial salient measurement which is defined as

$$D_i = \sum_{j=1}^{M} \left\| p_j^t - \mu_i^t \right\|^2 w(f_i^t, f_j^t), i = 1,2,3 \cdots M \tag{4}$$

where $w(f_i^t, f_j^t) = \frac{1}{Z_i^{'}} \exp(-\frac{1}{2\sigma_f^2} d(f_i^t, f_j^t))$ is also a Gaussian weight function. $\mu_i = \sum_{j=1}^{M} w(f_i^t, f_j^t) p_j$ defines the weighted mean position of super-pixel *i*. $Z_i^{'}$ is the normalization factor ensuring $\sum_{j=1}^{N} w(f_i^t, f_j^t) = 1$, and σ_f controls the range of the spatial salient operator.

The interpretation of Equation 4 is intuitive. Ideally features belonging to the background will be distributed over the entire image exhibiting a high spatial variance, whereas the object are generally more compact with small variance. For our spatial saliency, we slightly modify the ‘Element distribution’ proposed in [5] and instead computing in CIELab space with in L2ECM feature space to achieve a good performance in our experiments. In our implementation, parameters σ_p and σ_f are set to 5 and 15 respectively.

3. Energy Model for Salient Object Segmentation

Our salient object segmentation framework combines both temporal salient cue and spatial salient cue with object's appearance information. Based on the cues, segmentation can be solved by energy minimization.

Given input image *I*, let $\{I_i\}$ and $\{s_i\}$ denote the sets of image pixels and corresponding labels respectively. Label $s_i = 1$ if I_i belongs to the background, and $s_i = 0$ otherwise. $\{M_i^T\}$ and $\{M_i^D\}$ are the pixel-wise temporal salient map and the spatial salient map respectively (A regional salient value obtained from previous section is assigned to each pixel that belongs to the region). Salient object segmentation can be formalized as energy minimization

$$E(s) = \sum_{i \in I} \lambda_i^C U^C(s_i, I_i) + \sum_{i \in I} \lambda_i^T U^T(s_i, M_i^T) + \sum_{i \in I} \lambda_i^D U^D(s_i, M_i^D) + \sum_{(i,j) \in \Lambda} V(s_i, s_j, I_i, I_j) \tag{5}$$

where $U^C(s_i, I_i), U^T(s_i, M_i^T)$ and $U^D(s_i, M_i^D)$ are data association energies generated by object appearance cue, temporal salient cue and spatial salient cue respectively. λ_i^C, λ_i^T and λ_i^D are their weight factors for each pixel. They are all non-negative and satisfy $\lambda_i^C + \lambda_i^T + \lambda_i^D = 1$. $V(s_i, s_j, I_i, I_j)$ is a pairwise interaction energy between spatial neighboring pixels and Λ is a 4-connected neighbor system.

The interaction energy between I_i and I_j is defined as

$$V(s_i, s_j, I_i, I_j) = \delta(s_i \neq s_j) \frac{\varepsilon + e^{-\mu \left\| I_i - I_j \right\|^2}}{1 + \varepsilon} \tag{6}$$

where constant parameter $\varepsilon = 1$, μ_i is chosen to be $\frac{1}{2\left\langle \left\| I_i - I_j \right\|^2 \right\rangle}$, and $\langle \cdot \rangle$ denotes expectation over all pairs of neighbors in an image sample. This term imposes a tendency to spatial continuity of labels.

3.1. Object Appearance

Object appearance cue $U^C(s_i, I_i)$ evaluates the evidence for pixel labels based on color distribution in foreground and background. The foreground and background color likelihood are modeled non-parametrically according to the histograms in the YUV color space. The histograms is smoothed by a Gaussian filter to avoid over-learning, and they are learned adaptively over successive frames based on data from the segmented foreground in previous frame.

3.2. Temporal Saliency and Spatial Saliency

In addition to object appearance, temporal salient cue and spatial salient cue are key terms of our framework. Both cues are computed by the temporal salient map and the spatial salient map which are constructed in current frame. We define the temporal salient cue as

$$U^T(s_i, M_i^T) = \begin{cases} -\log \dfrac{\left|M_i^T - \mu^{T'}\right|}{\left|M_i^T - \mu^{T}\right| + \left|M_i^T - \mu^{T'}\right|} & s_i = 1 \\ -\log \dfrac{\left|M_i^T - \mu^{T}\right|}{\left|M_i^T - \mu^{T}\right| + \left|M_i^T - \mu^{T'}\right|} & s_i = 0 \end{cases} \quad (7)$$

where $\mu^{T'}$ and μ^{T} are the mean temporal salient value of the background and the one of the foreground respectively. They are updated by the previous segmentation result.

The spatial salient cue $U^D(s_i, M_i^D)$ is formalized in the same framework as $U^T(s_i, M_i^T)$, and only difference is to substitute $\{M_i^D\}$ for $\{M_i^T\}$. The object appearance cue, the temporal salient cue and the spatial salient cue are linearly combined with $\{\lambda_i^C\}, \{\lambda_i^T\}$ and $\{\lambda_i^D\}$ respectively. Such energy functions can be efficiently minimized by using the graph-cut algorithm as [2-1], leading to a binary segmentation of the image.

3.3. Online Weight Tuning

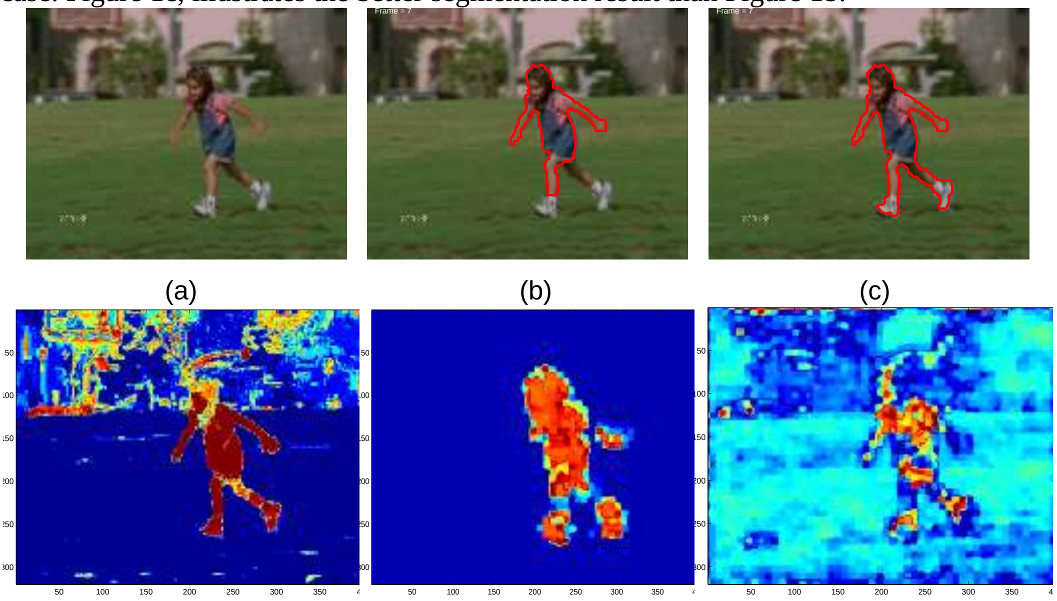
The weight factor of each energy item represents the proportion in the whole energy. Most proposed methods obtain the weight factors according to the global information of the corresponding cues. It is called *global weight* in our paper. Each weight is always computed by the discriminate performance of each segmentation cue. However, in some cases, each segmentation cue has different discriminate performances in different regions. The middle row of Figure 1, shows three segmentation cues in our paper. Overall, the temporal salient cue, Figure 1e, filters out most of the noise, and it will obtain the highest global weight in energy function. However, in terms of the object appearance cue, it has more ability to distinguish between the girl's legs and the background. Global weight will lead to inaccurate segmentation result which is illustrated in Figure 1b. *Local weight* for each pixel, which is computed by local information of each cue, is adopted in this case.

In our approach, we determine the discriminate performance λ_i of pixel *i* for each feature by using the pixels that lie inside a rectangle sub-window surrounding the pixel *i*. At each sub-window, two data clusters, $\left\{Z_F^{win}\right\}$ and $\left\{Z_B^{win}\right\}$ are extracted for each feature map respectively. $\left\{Z_F^{win}\right\}$ represents a data set of the feature value that belongs to the object in sub-window, and $\left\{Z_B^{win}\right\}$ is a data set that belongs to the background. A discriminate power of each pixel is measured in the sub-window which is defined as

$$\lambda_i = \max(0, \frac{\overline{Z_F^{win}} - \overline{Z_B^{win}}}{std(Z_F^{win}) + std(Z_B^{win})}) \quad s.t. \quad std(Z_F^{win}) \neq 0 \quad and \quad std(Z_F^{win}) \neq 0 \qquad (8)$$

where $\overline{Z_F^{win}}$ and $std(Z_F^{win})$ represent the mean and standard deviation of cluster $\left\{Z_F^{win}\right\}$ respectively. $\overline{Z_B^{win}}$ and $std(Z_B^{win})$ represent the mean and standard deviation of cluster $\left\{Z_B^{win}\right\}$. Because we find that the key factor affecting the segmentation result lies in the object's edge, the local weights are only computed in the neighborhood of the object's edge. And in other regions, global weight, computed by [2], is adopted as well.

The bottom row of Figure 1, shows three weighted segmentation cues. In Figure 1g, the features of the girl's legs are multiplied by larger weights to obtain the better discriminate power than global weight. In Figure 1h, the weighted features of the girl's head achieve the same effect. Although the saliency values in local weight (bottom row in Figure 1), are smaller than the one in global weight (middle row in Figure 1), on the whole, it can not affect the segmentation result because of the most parts of the girl are saliency in the image. In addition, the weight of the color likelihood in girl's legs and arms are larger than the weight of the other cues in the same locations. It will contribute to the accurate segmentation cut in graph-cut framework. The local weight solution definitely result in a better segmentation performance than global weight solution in this case. Figure 1c, illustrates the better segmentation result than Figure 1b.



(a) (b) (c)

(d) (e) (f)

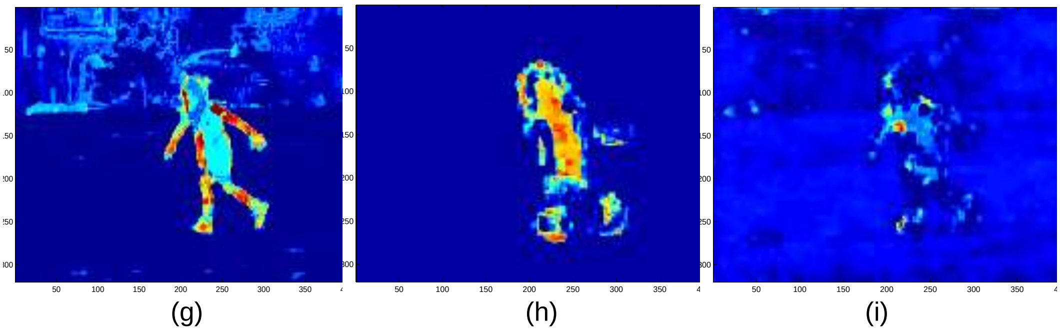


Figure 1(a): An Origin Image; (b): Segmentation Result by Global Weight; (c): Segmentation Result by Local Weight; (d): Color Likelihood; (e): Temporal Salient; (f): Spatial Salient; (g): Weighted Color Likelihood; (h): Weighted Temporal Salient; (i): Weighted Spatial Salient

4. Experiment

We use the GT-SegTrack database to perform quantitative performance comparisons between our approaches and two alternative segmentation methods: our global weight approach, our local weight approach, the state-of-the-art level set-based tracker described in [8] and the baseline graph-cut method using KLT-based temporal links described in [1]. Because the number of frames in each sequence is small and the scene in each sequence is not complicated, $\lambda_i^C, \lambda_i^T, \lambda_i^D$ are constantly set to 0.4, 0.4 and 0.2 respectively in our global weight approach. A quantitative comparisons of segmentation performances are provided in Table 1. The performance is measured by the average number of error pixels in each video sequence which is formalized as $\frac{\sum_{i=1}^{L} e_i}{L}$, where L is the number of the images in a sequence, and e_i is the number of error pixels in image i. The performances of our both approaches are better than other methods across some sequences. However, our global weight approach obtains a poor performance in girl sequence since the contours of the girl's legs and arms are drifted gradually. On the contrary, our local weight approach performs well in girl sequence because of the accurately segmentation in each frame. Figure 2, shows some examples of segmentation results based on our local weight approach in all sequences.

Table 1. Quantitative Comparison on GT-SegTrack Database

sequence	[8]	[1]	Our global weight	Our local weight	Average object size	Number of frames
parachute	502	**235**	300	298	3683	51
girl	1755	1304	3608	**1297**	8160	21
soldier	2984	2228	**1600**	1642	6321	31
monkey	4142	2814	2790	**2619**	6011	31

5. Conclusion

In this paper, we have proposed a novel energy minimization method for salient object segmentation. Temporal salient cue and spatial salient cue are computed by high dimensional Gaussian filtering function. Moreover, we deal with the weight of energy items in a generalized way. Adaptive pixel-wise weight approach is proposed to increase the robustness of the system in some cases. The approach is tested on several challenging video sequences and yields improved performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61271288) and National High Technology Research and Development Program 863(2012AA011503).

References

[1] D. Tsai, M. Flagg, A. Nakazawa and J. M. Rehg, “Motion coherent tracking using multi-label mrf optimization”, International journal of computer vision, vol. 100, no. 2, **(2012)**, pp. 190–202.

[2] Z. Yin and R. T. Collins, “Online figure-ground segmentation with edge pixel classification”, in BMVC. Citeseer, **(2008)**, pp. 1–10.

[3] W. Hu, W. Li, X. Zhang and S. Maybank, “Single and multiple object tracking using a multi-feature joint sparse representation”, Pattern Analysis and Machine Intelligence, vol. 37, no. 2, **(2015)**, pp. 816–833.

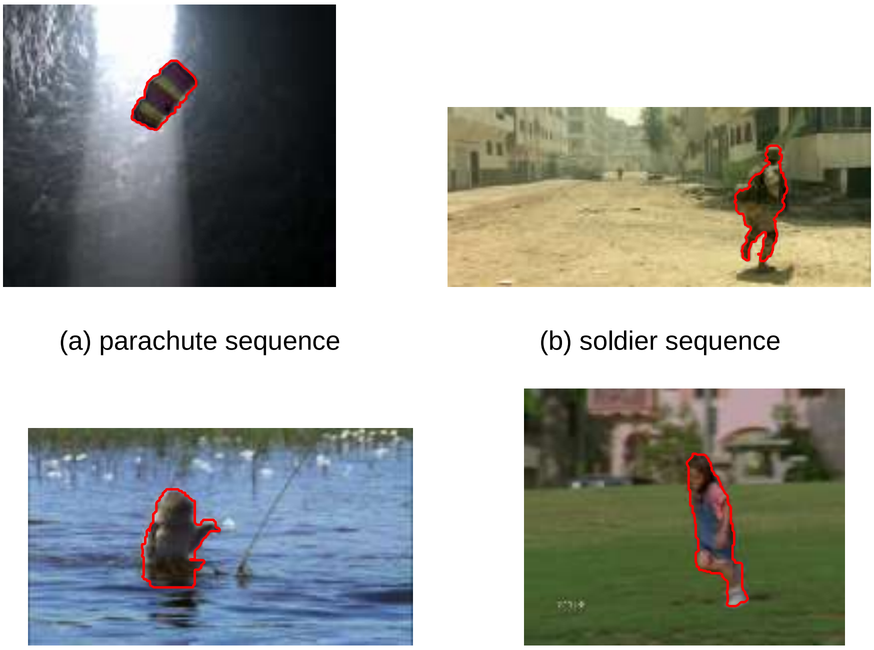
[4] L. Wang, M. Gong, C. Zhang, R. Yang, C. Zhang and Y. H. Yang, “Automatic real-time video matting using time-of-flight camera and multichannel poisson equations”, International journal of computer vision, vol. 97, no. 1, **(2012)**, pp. 104–121.

[5] F. Perazzi, P. Krahenbuhl, Y. Pritch and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection”, in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, **(2012)**, pp. 733–740.

[6] P. Li and Q. Wang, “Local log-euclidean covariance matrix (l2ecm) for image representation and its applications,” in Computer Vision–ECCV 2012. Springer, **(2012)**, pp. 469–482.

[7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Su sstrunk, “Slicsuperpixels”, Tech.Rep., **(2010)**.

[8] P. Chockalingam, N. Pradeep and S. Birchfield, “Adaptive fragments-based tracking of non-rigid objects using level sets”, in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, **(2009)**, pp. 1530–1537.



(a) parachute sequence

(b) soldier sequence

(c) monkey sequence

(d) girl sequence

Figure 2. Some Examples of Segmentation Results Based on Our Local Weight Approach